PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	:	
Mark D. WASSON et al.	:	Confirmation No.: 4117
Serial No. 10/716,202	:	Art Unit: 3624
Filed: November 19, 2003	:	Examiner: (not yet assigned)
For: EXTRACTION OF FACTS FROM TEXT	:	Atty Docket: P68795US0

INFORMATION DISCLOSURE STATEMENT UNDER 37 CFR 1.97

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith on form PTO-1449 is a listing of documents known to Applicants and/or their attorney in compliance with the requirements of 37 CFR 1.56. Copies of the documents are also being submitted.

Since the Information Disclosure Statement is filed before the issuance of the first Office Action on the merits or within three months of filing, no fee is required. The Commissioner is hereby authorized to charge fees under 37 CFR § 1.17 which may be required for consideration of this paper; or credit any overpayment to Deposit Account No. 06-1358.

In accordance with 37 CFR § 1.97(g) and (h), the filing of this Information Disclosure Statement shall not be construed as a representation that a search has been made, or as an admission that the information cited herein is, or is considered to be, material to patentability as defined in 37 CFR § 1.56(b).

Applicants do not waive any rights to appropriate action to establish patentability over the listed documents should they be applied as references against the claims of the present application.

RELEVANCE OF THE DOCUMENTS

Documents AF-AG relate to the subject of information extraction, generally. Documents AH-CG relate to the subject of pattern recognition languages. Documents AM and BF also relate to the subject of uncrossing (handling annotation conflicts).

1. Document CG

B. Baldwin, "EAGLE: An Extensible Architecture for General Linguistic Engineering," *Proceedings of RIAO '97* (June 1997), pp. 271-283):

In 1996, LexisNexis funded a collaborative research project at the University of Pennsylvania that resulted in a proof of concept fact extraction system prototype that served as a starting point and inspiration for its subsequent work on fact extraction. Dubbed "EAGLE" in Document CG and the "Penn Tools" in LexisNexis-internal documentation, this prototype was in many ways typical of the state of the art of fact extraction systems at that time. It used several linguistic and other annotation processes to assign attributes to text, and it used a regular expression-based pattern recognition

language called Mother of Perl (MOP) to recognized patterns of attributes that correspond to text in some document to be extracted.

In the Penn Tools, annotations results were stored in separate but aligned "analytical tiers", referred to as "a parallel file data structure" (p. 274). Because annotations were stored separately in position-based files rather than in trees, there was no need to resolve conflicts between annotations, and there were no tree-based representations to navigate during pattern recognition. XML in fact did not exist when the Penn Tools (EAGLE) prototype was created, so that it did not require addressing the problems resulting from the decision in the present invention to use XML instead of the analytical tiers model.

Although document CG paper presents information on a historical prototype that contributed to the creation of the present invention, it does not describe the uncrossing approach to eliminate conflicts between annotations to create a well-formed XML representation of the annotations, and the combination of regular expressions and tree navigation in the pattern recognition language, which are important components of the present invention.

2. Documents Providing Background Information on Fact Extraction

a. Document AF

R. Grishman et al., "Message Understanding Conference - 6: A Brief History" (Nov. 1996), *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, (June 1996), pp. 466-471:

The U.S. defense and intelligence community sponsored a series of conferences called Message Understanding Conferences to promote research into fact extraction and related technologies. Participants would all be provided the same problem definition and test documents. Participants then created and tested fact extraction applications to solve the specified problem on the provided documents. Participants then met compare results and to share their experiences and lessons learned.

Document AF provides general information on the MUC series, and on the sixth conference in this series. Two of the inventors of the present application, Mark Wasson and Valentina Templar, participated in MUC-6.

Document AF does not describe any of the systems presented at MUC-6, and because MUC-6 predated the creation of XML, none of the participating systems had to address the problems addressed in the present invention Set that resulted from the use XML. Thus document AF is not directly relevant to the present invention, although it provides the reader with background information on general research conducted in the fact extraction area.

b. Document AG

D. Appelt et al., "Introduction to Information Extraction Technology, A Tutorial Prepared for IJCAI-99," *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (July 31 - August 6, 1999), pp. 1-41:

These tutorial notes provide the reader with an introductory overview of the general problem of fact extraction. Because it serves as an introduction to the topic, it provides discussion on components common to a number of fact extraction tool sets, including that of the present invention.

These include the text processing components, or annotators, listed in The Architecture of Information Systems (p. 10) and The Components of an Information Extraction System (p. 13) sections.

The Extraction of Domain-specific Relationships and Events (p. 29) section contrasts rule-based approaches to fact extraction, like the present invention, to machine learning-based approaches, but it does not describe specific rule-based pattern recognition languages.

The tutorial does not focus on a particular system although it does use some examples based on the authors' experience with their own FASTUS extraction tool set.

3. Documents Focused on Pattern Recognition Language Functionality

One novel aspect of the present invention is its combination of regular expression functionality with tree traversal functional as a means for writing rules for recognizing patterns of annotated features and syntactic relationships that correspond to some text to be extracted from a document. This section surveys documents that describe how other implementations perform pattern recognition. Typically, other fact extraction systems use either a regular expression-based pattern recognition language to identify and extract text, or they used statistical machine learning.

a. Documents relating to GATE

The GATE (General Architecture for Text Engineering) fact extraction tool kit was developed at the University of Sheffield, and today it is one of the better known fact extraction tool kits.

i. **Document AJ**

H. Cunningham, "Software Architecture for Language Engineering," Ph.D. Thesis, Department of Computer Science, University of Sheffield (June 2000), Chapter 7 and Appendix A:

Chapter 7 (p. 127) describes the design and implementation of GATE. Appendix A Java Annotation Patterns Engine Specification (p. 185) provides a formal description of the pattern recognition language used in GATE. The pattern recognition language is regular expression-based (p. 185), and is not combined with tree navigation as is done with RuBIE in the present invention.

ii. **Document AK**

H. Cunningham et al., "Experience of using GATE for NLP R&D," *Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*, (Luxembourg 2000), pp. 1-8:

This paper provides a basic overview of GATE. It says little about its architecture or pattern recognition process, and thus is provided as background information on that system.

iii. **Document BF**

H. Cunningham, "Developing Language Processing Components with GATE (a User Guide) For GATE version 2.1 beta 1 (August 2002)," University of Sheffield (2001-2002):

This document provides guidance for GATE users. In their discussion on how it represents annotations (Section 5.4), the authors report using directed acyclic graphs. Chapter 6 reports once again that GATE uses regular expressions for pattern recognition.

iv. **Document BI**

H. Cunningham et al., "GATE: an Architecture for Development of robust HLT Applications," *Proceedings of ACL 2002* (2002), pp. 1-8:

This paper provides an introductory overview of GATE and is provided as background information.

v. **Document BJ**

D. Maynard, "Architectural elements of Language Engineering Robustness," *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data* 1 (1):1-20 (2002):

This paper provides an overview of the system architecture that underlies GATE, and is provided as background information.

vi. **Document BG**

K. Bontcheva, "Using Human Language Technology for automatic Annotation and Indexing of Digital Library Content," *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries* (2002), pp. 613-625:

This paper describes an example GATE-based application and is provided as background.

b. **Document relating to Brightware**

The MITA system applies information extraction techniques to Met Life insurance application content.

i. **Document AH**

B. Glasgow, "MITA: An Information Extraction Approach to Analysis of Free-form Text in Life Insurance Applications," *AI Magazine*, 19(1):59-71, 1998:

This paper describes MITA, which uses ART*Enterprise technology as a basis for its information extraction. MITA parses text fragments, but uses a rule-based system that relies on IF-THEN-style statements as a basis for selecting and extracting targeted information insurance application fields.

c. **Document relating to ClearForest.**

ClearForest uses regular expression-like functions to support its pattern recognition.

i. **Document AI**

R. Feldman et al., "Text Mining at the Term Level," *Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery* (Nantes, France, Sept 1998), pp. 1-9:

This paper provides information on early work at ClearForest. In subsequent contacts with ClearForest, the example pattern recognition rules the inventors have seen have been lists of function-like components, which could be regarded as a low-level regular expression. ClearForest does hold a patent on its work, U.S. Patent No. 6,442,545, entitled "Term-level Text with Mining with Taxonomies," which is submitted herewith as Document AA.

ii. **Document BH**

R. Feldman, "Mining biomedical literature using information extraction," *Current Drug Discovery* (October 2002) pp. 19-23:

This paper provides an overview of the current status of ClearForest's approach to information extraction, including a description of how it uses limited parser information and an example rule written in DIAL, their pattern recognition language.

d. Document relating to New York University's Proteus system

NYU's Proteus system uses a finite state transducer as the basis for its pattern recognition.

i. Document AL

R. Grishman, "Real-Time Event Extraction for Infectious Disease Outbreaks," *Proceedings of Human Language Technology Conference (HLT)* (2002), pp. 1-4:

This paper describes an application built using Proteus. Section 5. Extraction Engine reports their use of a finite state transducer.

e. Document relating to SRI FASTUS

SRI uses cascaded finite state automate, effectively regular expressions applied to layers of annotations, to support its pattern recognition. (p. 6)

i. Document BK

J. Hobbs et al., "FASTUS: Extracting Information from Natural-Language Texts," *Finite State Devices for Natural Language Processing* (MIT Press 2000), pp.1-22:

This paper provides an overview of the fact extraction problem and the FASTUS system that SRI has created to address it.

f. Document relating to BBN

IdentiFinder is a commercialized version of BBN's fact extraction technology.

i. **Document BL**

This paper describes BBN's approach to fact extraction. BBN combines general linguistic feature annotation and parsing. Because their annotators were designed and integrated to work together, they do not have conflicts in their annotations and thus do not have the uncrossing problem. For pattern matching, instead of a rule-based approach, BBN uses statistical machine learning and thus does not have the combination of rule-based regular expression pattern matching and tree navigation found in the present invention.

g. **Document relating to Hapax**

The Hapax FindEngine™ is not a fact extraction system in any traditional sense. Instead, it identifies linguistically-based relationships in text, such as subject-verb-object relationships. User then search this collection of extracted facts. It is more appropriate to regard this as question answering technology than fact extraction technology. However, a collection of questions about some common topic could retrieve a relevant collection of extracted relationship facts about that topic. Hapax appropriately refers to this as fact retrieval instead of fact extraction.

i. **Document CF**

FindEngine™ white paper, Version 1.0, Hapax Information Systems AB (September 2001) pp. 1-14:

This paper describes the Hapax model. It provides a view of an information retrieval-based alternative to fact extraction and thus is provided as background information.

4. Documents Focused on Handling Annotation Conflicts

Annotation conflicts occur when the scope of text covered by two annotations both partially overlaps and partially does not overlap. In the example

```
ABCDEFGHIJ
  KKKKK
  LLLL
    MM
       NNNNN
     OOO
```

In this example OOO conflicts with KKKKK, LLLL, MM and NNNNN because for each pair each annotation spans a part of the text ABCDEFGHIJ that the other does not. KKKKK and LLLL are not in conflict because all of LLLL is covered by the scope of KKKKK. KKKKK and NNNNN are not in conflict because their scopes do not overlap at all.

Conflicts are generally not a problem except when the annotations are stored in a hierarchical representation, and in particular one that required well-formedness. XML requires well-formedness.

a. Document AM

B. Crysmann, "An Integrated Architecture for Shallow and Deep Processing," *Proceedings of ACL-2002, Association for Computational Linguistics 40th Anniversary Meeting* (July 2002), pp. 1-8:

The inventors have found only one system whose creators rely on XML to represent the annotations and who address the problem of conflicting annotations, the WHITEBOARD system created at the German Research Center for Artificial Intelligence (DFKI). However, DFKI separates different annotations into separate annotation layers (first paragraph, Section 2, Architecture.) By

representing potentially conflicting annotations separately, DFKI has no need to uncross annotations to create a single well-formed XML representation.

b. Document BF

H. Cunningham, "Developing Language Processing Components with GATE (a User Guide) For GATE version 2.1 beta 1 (August 2002)," University of Sheffield (2001-2002):

This document is also discussed in the pattern recognition section (section 3, page 4, above). Although GATE uses directed acyclic graphs rather than XML to represent its annotations, it can apply to input documents in XML format, and it can produce output documents in XML format. However, the authors do note in Section 5.5.2 XML Subsection Output:

> In order to understand why there are two types of XML serialization, one needs to understand the structure of a GATE document. GATE allows a graph of annotations that refer to parts of the text. Those annotations are grouped under annotation sets. Because of this structure, sometimes it is impossible to save a document as XML using tags that surround the text referred by the annotation, because tags crossover situations could appear (XML is essentially a tree-based model of information, whereas GATE uses graphs). Therefore, in order to preserve all annotations in a GATE document, a custom type of XML document was developed.
> The problem of crossover tags appears with GATE's second option (the preserve format one), which is implemented at the cost of losing certain annotations. The way it is applied in GATE is that it tries to restore the original markup and where it is possible, to add in the same manner annotations produced by GATE.

Thus noting the problems that crossover tags cause for them when trying to output documents in XML. Rather than deploy something analogous to the uncrossing algorithm of the present invention, GATE loses the problem annotations.

CONCLUSION

It is respectfully requested that the Examiner initial and return a copy of the enclosed Form PTO-1449, and to similarly indicate in the official file wrapper of this patent application that the attached documents have been considered.

If the Examiner has any questions or wishes to discuss this application, the Examiner is invited to telephone the undersigned representative at the number set forth below.

Respectfully submitted,

JACOBSON HOLMAN PLLC

Date: Mar. 9, 2004

By: _____

Customer No. 00,136
400 Seventh Street, N.W.
Washington, D.C. 20004
(202) 638-6666

Linda J. Shapiro
Registration No. 28,264

Form PTO-1449

Information Disclosure Citation

Attorney Docket	Application No.
P68795US0	10/716,202
Applicant: Mark D. WASSON et al.	
Filing Date: November 19, 2003	Group Art Unit:

U.S. Patent Documents

Examiner Initial		Patent Number	Date	Name	Class	Sub-Class	Filing Date
	AA	6,442,545	08/27/2002	Feldman et al.			
	AB						
	AC						

Foreign Patent Documents

		Document Number	Date	Country	Class	Sub-Class	Trans-lation
	AD						Yes No
	AE						Yes No

Other Documents (Including Author, Title, Date, Pertinent Pages, etc.)

	AF	R. Grishman et al., "Message Understanding Conference - 6: A Brief History" (Nov. 1996), *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, (June 1996), pp. 466-471.
	AG	D. Appelt et al., "Introduction to Information Extraction Technology, A Tutorial Prepared for IJCAI-99," *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (July 31 - August 6, 1999), pp. 1-41.
	AH	B. Glasgow, "MITA: An Information Extraction Approach to Analysis of Free-form Text in Life Insurance Applications," *AI Magazine*, 19(1):59-71, 1998.
	AI	R. Feldman et al., "Text Mining at the Term Level," *Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery* (Nantes, France, Sept 1998), pp. 1-9.
	AJ	H. Cunningham, "Software Architecture for Language Engineering," Ph.D. Thesis, Department of Computer Science, University of Sheffield (June 2000), p. i (Abstract), Table of Contents, List of Figures, Chapter 7, and Appendix A.
	AK	H. Cunningham et al., "Experience of using GATE for NLP R&D," *Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*, (Luxembourg 2000), pp. 1-8.
	AL	R. Grishman, "Real-Time Event Extraction for Infectious Disease Outbreaks," *Proceedings of Human Language Technology Conference (HLT) (2002)*, pp. 1-4.
	AM	B. Crysmann, "An Integrated Architecture for Shallow and Deep Processing," Proceedings of ACL-2002, *Association for Computational Linguistics 40th Anniversary Meeting* (July 2002), pp. 1-8.

Examiner	Date Considered

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP § 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

Form PTO-1449

Information Disclosure Citation

Attorney Docket	Application No.
P68795US0	10/716,202
Applicant: Mark D. WASSON et al.	
Filing Date: November 19, 2003	Group Art Unit:

U.S. Patent Documents

Examiner Initial		Patent Number	Date	Name	Class	Sub-Class	Filing Date
	BA						
	BB						
	BC						

Foreign Patent Documents

		Document Number	Date	Country	Class	Sub-Class	Trans-lation
	BD						Yes No
	BE						Yes No

Other Documents (Including Author, Title, Date, Pertinent Pages, etc.)

	BF	H. Cunningham, "Developing Language Processing Components with GATE (a User Guide) For GATE version 2.1 beta 1 (August 2002)," University of Sheffield (2001-2002) Table of Contents, and Chapters 5 and 6.
	BG	K. Bontcheva, "Using Human Language Technology for automatic Annotation and Indexing of Digital Library Content," *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries* (2002), pp. 613-625.
	BH	R. Feldman, "Mining biomedical literature using information extraction," *Current Drug Discovery* (October 2002) pp. 19-23.
	BI	H. Cunningham et al., "GATE: an Architecture for Development of robust HLT Applications," *Proceedings of ACL 2002* (2002), pp. 1-8.
	BJ	D. Maynard, "Architectural Elements of Language Engineering Robustness," *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data* 1 (1):1-20 (2002).
	BK	J. Hobbs et al., "FASTUS: Extracting Information from Natural-Language Texts," *Finite State Devices for Natural Language Processing* (MIT Press 2000), pp.1-22.
	BL	S. Miller et al., "A Novel Use of Statistical Parsing to Extract Information from Text," *6th Applied Natural Language Processing Conference* (2000), pp. 1-8.
	BM	J. Hobbs et al., "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text," *Finite State Devices for Natural Language Processing* (MIT Press 1996), pp.1-22.

Examiner	Date Considered

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP § 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

Form PTO-1449

Information Disclosure Citation

Attorney Docket	Application No.
P68795US0	10/716,202
Applicant: Mark D. WASSON et al.	
Filing Date: November 19, 2003	Group Art Unit:

U.S. Patent Documents

Examiner Initial		Patent Number	Date	Name	Class	Sub-Class	Filing Date
	CA						
	CB						
	CC						

Foreign Patent Documents

		Document Number	Date	Country	Class	Sub-Class	Trans-lation
	CD						Yes No
	CE						Yes No

Other Documents (Including Author, Title, Date, Pertinent Pages, etc.)

	CF	*FindEngine™ white paper, Version 1.0*, Hapax Information Systems AB (September 2001) pp. 1-14.
	CG	B. Baldwin, "EAGLE: An Extensible Architecture for General Linguistic Engineering," *Proceedings of RIAO '97* (June 1997), pp. 271-283.
	CH	
	CI	
	CJ	
	CK	
	CL	
	CM	

Examiner	Date Considered

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP § 609. Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.